一种改进的个性化查询引文推荐方法 *

李 飞[1]，张宏鸣[1†]，蔡晓妍[2]，刘 斌[1]，郭蓝天[2]

(1. 西北农林科技大学 信息工程学院，陕西 杨陵 712100; 2. 西北工业大学 自动化学院，西安 710072)

摘 要：为充分利用文本内容的上下文信息，结合图模型及查询向量的构建方法，提出一种融合查询内容信息的个性化引文推荐方法。通过三种论文信息构建三层图模型，并在不同层上设置不同参数，调整节点向不同层次的跳转概率；利用 word2vec 技术构建的查询向量，可以有效利用文本上下文内容信息，使相似的文章在距离上更加接近，进而对候选文章进行评分预测与论文推荐。在 Association of Computational Linguistics Anthology Network 数据集上进行计算分析，相同查询下与原有的方法相比在 recall@N 上平均提高约 7%，在 NDCG@N 上平均提高约 11%。实验结果表明该方法可以使引文推荐的质量得到有效的提升，能够获得较好的推荐效果。

关键词：多关系图；词向量；查询向量；带重启的随机游走；个性化推荐

中图分类号：TP301.6 doi: 10.3969/j.issn.1001-3695.2018.02.0086

Improved method for personalized query citation recommendation

Li Fei[1], Zhang Hongming[1†], Cai Xiaoyan[2], Liu Bin[1], Guo Lantian[2]

(1. *College of Information Engineering*, *Northwest A & F University*, *Yangling Shanxi* 712100, *China*; 2, *School of Automation*, *Northwestern Polytechnical University*, *Xi'an* 710072, *China*)

Abstract: To make full use of the context information of the papers, combined with the construction method of graph model and query vector, this paper proposed a fusion query information personalized citation recommendation method. Built a three layer graph model through three kinds of paper information, and set different parameters on different layers to adjust the jump probability of nodes to different levels; the query vector constructed using word2vec technology can effectively use the text context information, so that similar papers are closer to the distance, and then the candidate papers are predicted and recommended. Computational analyzes performed on the Association of Computational Linguistics Anthology Network dataset showed an average increase of about 7% over recall@N and an average increase of about 11% over NDCG@N for the same query compared to the original method. Experimental results show that the proposed method can effectively improve the quality of citation recommendation and get better recommendation results.

Key words: multi-relation; word vector; query vector; random walk with restarts; personalized recommendation

0 引言

科技论文推荐的研究是为满足研究人员引用需求，推荐少量的并与他们研究内容相关的科技论文[1]。目前，使用最广泛的推荐技术是基于内容过滤（content-based filtering，CBF）和基于协同过滤（collaborative filtering，CF）[2]。CBF 从文本内容中提取到的单词或者根据一定方法获得文章主题层面的研究内容，从而推荐与查询相匹配的论文。但同时 CBF 具有传统信息检索所存在的问题，如语义模糊性[3]。CF 方法主要是利用论文引用网络和作者合作关系网络来挖掘论文和查询者之间相互关系进行推荐，用到的信息如用户标签[4]或用户历史信息[5]，但 CF 这种方法具有数据稀疏、冷启动和可扩展性等问题[6,7]。

随着复杂异构信息网络研究的兴起，基于图模型的推荐方法受到越来越多的关注[8]。现有基于图的推荐方法使用数据集各种关联信息等来构建图模型（如文献[9]），把引文推荐作为一项引用链接预测工作，如 West 等人[10]提出的基于引用关系图的层次聚类算法来确定论文的相关性，并根据论文在这些聚类中的重要性进行推荐。为解决引用关系图的稀疏和噪音问题，

收稿日期：2018-02-09；**修回日期：**2018-03-21 **基金项目：**国家自然科学基金资助项目（41771315，41301283，41371274）；国家重点研发计划资助项目（2017YFC0403203）；欧盟地平线 2020 研究与创新计划资助项目（GA：635750）；陕西省自然科学基金面上项目（2017JM6059）

作者简介：李飞（1989-），男，河南泌阳人，硕士研究生，主要研究方向为基于深度学习的文本推荐；张宏鸣（1979-），男（通信作者），内蒙古赤峰人，副教授，硕导，主要研究方向为空间大数据管理与区域土壤侵蚀评价研究（zhm@nwsuaf.edu.cn）；蔡晓妍（1982-），女，山东淄博人，副教授，主要研究方向为是文本摘要、信息检索、机器学习；刘斌：（1981-），男，陕西渭南人，博士研究生，主要研究方向为是并行计算、机器学习；郭蓝天（1987-），博士研究生，主要研究方向为数据挖掘及机器学习等．

Zhou 等人[11]提出了一种新的整合了引用关系和作者关系来衡量论文相似度的推荐任务的分解策略图模型。然而，该方法主要强调引文网络的链接作用，忽视了数据集中的其他有用信息。Pan 等人[9]利用引用关系和内容信息形成一个异构网络，使用基于图的相似度学习算法执行推荐任务。Meng 等人[12]基于作者，论文，词和主题四个层面构建异构图模型，提出了个性化推荐算法，在构建关键字层时用到了 latent Dirichlet allocation（LDA）[13]或者 TF-IDF[14,3]。但是 TF-IDF 只是词之间的重叠，没有考虑词语之间的语义信息。Ren 等[1]认为一些目标文章查询进行推荐结果产生的过程是不同的，因此他们提出了一种软聚类方法来解释这种行为差异。Wang 等人[15]通过基于主题建模和内容分析的概率模型将文本内容融入到传统的矩阵分解方法中去。

上述方法存在两个方面的问题：a）这些方法没有考虑没有考虑到词与词之间的上下文之间的关系；b）查询中没有用到查询内容与候选集合论文的关系。本文利用作者关系、论文引用关系和论文内容等信息构建三层图模型，在层与层之间添加相关参数来控制模型中节点跳转的不确定性。在此基础上，利用研究人员的个性化信息和上下文内容信息构建查询向量向量，运行带重启的 Random Walk 的方法，产生最终的推荐列表。本文所提方法具有以下贡献：a)文本内容表示向量的生成，可以更有效的利用文本上下文信息；b)在不同节点层之间设置不同的游走参数，可以使节点的跳转更符合实际情况，能够对目标的推荐结果进行优化。

1 本文方法

1.1 问题定义

为了能够简单的表示和直观的理解，在本文中把文献推荐问题定义为训练得到相关论文评分列表的问题 $\boldsymbol{r}(\boldsymbol{q}, \boldsymbol{p}):\boldsymbol{Q}\times\boldsymbol{P}\rightarrow\boldsymbol{R}$，其中 $\boldsymbol{q}$ 表示针对一篇文章的一个查询向量，$\boldsymbol{q}\in\boldsymbol{Q}$；$\boldsymbol{p}$ 为一篇候选文献，$\boldsymbol{p}\in\boldsymbol{P}$；$\boldsymbol{R}$ 是查询结果集合，此问题是根据论文集合中的异构关系构成的。通过训练得到的针对查询文献与候选文献之间的得分列表 $\boldsymbol{r}(\boldsymbol{q}, \boldsymbol{p})$，然后根据 $\boldsymbol{r}(\boldsymbol{q}, \boldsymbol{p})$产生推荐结果集合。更详尽地，本文把个性化引文推荐问题定义如下：给定一个由论文集合中相关信息构建的异构图 G 和查询论文 $\boldsymbol{q_p}$ 查询关键字 $\boldsymbol{q_w}$ 和查询人 $\boldsymbol{q_a}$（如果该查询者是已知的）组成一个查询向量 $\boldsymbol{q}=[\boldsymbol{q_p}, \boldsymbol{q_a}, \boldsymbol{q_w}]$。根据构建的推荐模型，为一个查询 $\boldsymbol{q}$ 推荐一个与目标文献相关的候选文献子集。表 1 是给出了本文用到的一些标记符号。

表 1 符号标记

标记符	标记符描述
$\boldsymbol{P}$	论文集合
$\boldsymbol{A}$	作者集合
$\boldsymbol{W}$	关键字集合
$\boldsymbol{Q}$	查询向量集合
m	作者数量
n	论文数量
k	关键字个数
$\boldsymbol{M_{pp}}$	$n\times n$ 的论文引用关系矩阵
$\boldsymbol{M_{aa}}$	$m\times m$ 的作者关系矩阵
$\boldsymbol{M_{ww}}$	$k\times k$ 的关键字矩阵
$\boldsymbol{Map}$	$m\times n$ 的作者-论文关系矩阵
$\boldsymbol{Mpw}$	$n\times k$ 的论文-关键字关系矩阵
$\boldsymbol{q}$	$n+m+k$ 个元素的查询向量
$\boldsymbol{q_p}$	包含于 $\boldsymbol{q}$ 的 n 个元素的论文向量
$\boldsymbol{q_a}$	包含于 $\boldsymbol{q}$ 的 m 个元素的作者向量
$\boldsymbol{q_w}$	包含于 $\boldsymbol{q}$ 的 m 个元素的关键字向量
θ	词向量相似度阈值

1.2 基于个性化查询的文献推荐方法

1.2.1 三层图模型

为了尽可能有效地利用数据集中信息进行推荐，本文构建了多层图网络模型。如图 1 所示，该图模型包含三种不同类型的实体—作者、论文、关键字，存在的关系有作者-作者合作关系（R_1），作者-论文关系（R_2），论文-论文引用关系（R_3），论文-关键字包含关系（R_4）等四种。因此，它们之间的关系可以归结为一个三层图模型，层内之间连接的两点表示两个相同的实体类型，层间的链接的两点表示两个不同实体类型。该三层图模型可以表示如下：$G=<V,E,M>$,其中 V 表示的是节点的集合。V 包含三个集合：作者集合 $A=\{a_1,a_2,\ldots,a_m\}$，论文集合 $P=\{p_1,p_2,\ldots,p_n\}$，关键字集合 $V=\{w1,w2,\ldots,wk\}$,也即 $V=A\cup P\cup V$。集合 E 是集合 V 中存在链接的边，$E=\{<vi,vj>|\ vi,vj\in V\}$。$M$ 是关系矩阵，其中 wij 表示链接节点 vi 和节点 vj 之间的权重。

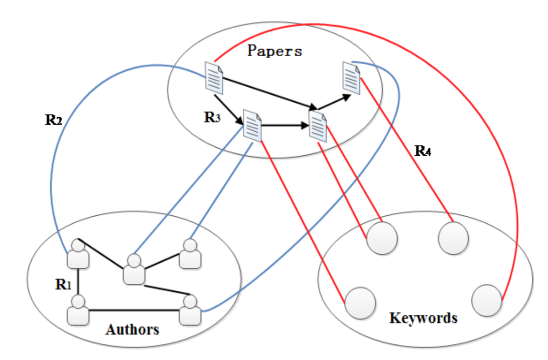


图 1 三层图模型

a)作者层。当查询相关文档时，作者的可靠性和专业知识可以发挥重要的作用[16]。在构建作者关系矩阵 $\boldsymbol{M_{AA}}$ 时，如果作者 a_i 和 a_j 有合作关系，则 $w_{ij}=w_{ji}=1$；如果两个作者没有合作关系 $w_{ij}=w_{ji}=0$。同理，在 $\boldsymbol{M_{AP}}$ 中，如果作者 a_i 是论文 p_j 的一个作者，则 $w_{ij}=w_{ji}=1$；如果不是则置 $w_{ij}=w_{ji}=0$。

b)关键字层。在 $\boldsymbol{M_{WW}}$ 中设置所有边权值为 0，对于 M_{PW} 和上文作者与论文关系类似，如果论文 p_i 中包含关键字 v_j，$w_{ji}=w_{ij}=1$，否则 $w_{ji}=w_{ij}=0$。

c)论文层。与上面不同，在 $\boldsymbol{M_{PP}}$ 中，如果论文 p_i 引用了论文 p_j，$w_{ij}=1$，则 $w_{ji}=0$；如果两篇论文之间不存在引用关系则

$w_{ij} = w_{ji} = 0$。

根据上文所述，本文中概率转移矩阵 **M** 被分解为九个子矩阵，即 M_{AA}，M_{PP}，M_{WW}，M_{AP}，M_{AW}，M_{PW}，M_{WP}，M_{PA}，M_{WA}，其中 $M_{AW}=0$，$M_{WA}=0$。一篇候选文章与目标文献的相关性可以通过层加权和得到，计算如下所示：

$$R(p_i)=(1-\theta)\left[\alpha\sum_{p_j\to p_i}M_{PP}^{ji}R(p_j)+\beta\sum_{a_j\to p_i}M_{PA}^{ji}R(a_j)+\gamma\sum_{w_j\to p_i}M_{PW}^{ji}R(w_j)\right]+\theta q \quad (1)$$

其他层节点的相关性可以通过相似的公式计算得到。从式（1）可知通过调整相应的参数（α，β，γ）可以控制层节点从当前层游走的下一层的概率。分解后的关系矩阵 **M** 可以表示成如下所示：

$$M=\begin{pmatrix}\alpha_1 M_{PP} & \beta_1 M_{PA} & \gamma M_{PW}\\ \alpha_2 M_{AP} & \beta_2 M_{AA} & 0\\ \alpha_3 M_{WP} & 0 & 0\end{pmatrix}$$

在分解矩阵 **M** 中，$M_{PP}(i,j)$ 表示三层图模型中论文 p_i 和 p_j 之间的引用关系矩阵，$M_{AA}(i,j)$ 表示模型中的作者是作者 a_i 和作者 a_j 的合作关系矩阵，$M_{AP}(i,j)$ 表示模型中的作者-论文关系矩阵，$M_{PW}(i,j)$ 表示图模型中论文-关键字关系矩阵。并且在矩阵 **M** 中 $M_{AP}=M_{PA}^{T}$，$M_{WP}=M_{PW}^{T}$。各个位置所对应的值如上文作者层，关键字层，论文层所述。

在该模型中，在论文层的一个节点，有三种可能的运动行为：运动到作者层或者是关键字层，亦或者是还在论文层，所以本文设定 $\alpha_1+\alpha_2+\alpha_3=1$。本文还假定 $\beta_1+\beta_2=1$，因为一个节点再次游走之后及可能在作者层，也有可能调出作者层。类比于关键字层，设 $\gamma=1$。通过这些约束简化参数个数，所以矩阵 **M** 可以有如下表示：

$$M=\begin{pmatrix}\alpha_1 M_{PP} & \beta M_{PA} & M_{PW}\\ \alpha_2 M_{AP} & (1-\beta)M_{AA} & 0\\ (1-\alpha_1-\alpha_2)M_{WP} & 0 & 0\end{pmatrix}$$

1.2.2 查询向量

与 Totti 等人[3]的查询向量不同，本文所采用的三层模型的查询向量 q 的构成如下所示：

a)针对作者的查询 q_a 表示查询作者已经确定，针对一个查询对象 u，如果该作者存在于候选作者集合中，则有 $q_a(u)=1$，否则，$q_a(i)=0$，$i\neq u$。

b)在进行查询时，研究人员输入的查询内容是几个独立的词汇，而非具有完整语义信息的文本，利用文本相似度计算方法难以得到。词向量化 word embedding[17]可以将词映射为特征向量，利用向量之间的距离来逼近词与词之间的语义。因此本文采用 word2vec 中的 CBOW 词向量化模型[17]生成词向量。设研究人员输入的原始查询 q_r 为 m 个单词的集合，即 $q_r=\{w_i|i=1,2,\ldots,m\}$。利用词向量模型训练语料库，可以得到 q_r 中各个单词的 N 维向量，单词 w_i 表示为 $w_i=\left[w_i^1,\ \cdots\ w_i^N\right]$，文章向量本文用 q_r 中各个单词的词向量进行累加得到，得到 N 维的文章内容表示。计算公式如下所示：

$$q_r=\left[\sum_{i=1}^{n}w_i^1,\ \sum_{i=1}^{n}w_i^2,\ \cdots,\ \sum_{i=1}^{n}w_i^{N-1},\ \sum_{i=1}^{n}w_i^N\right] \quad (2)$$

其中:n 为某篇文章的单词数量。同样的方法可以得到图模型中各个文章的内容进行 N 维向量空间的表示，记为 $pv=\{pv_j|j=1,2,\ldots,k\}$，其中 k 为图模型中的文章数量。若查询向量 $q_r=X=[x_1,\ldots,x_N]$，候选文章向量 $pv=Y=[y_1,\ldots,y_N]$，通过 consine 相似度计算 q_r 内容向量与各个候选文章 pv 的内容向量之间的相似度得 $consine(q_r, pv)$=consine（X,Y），计算公式如下所示：

$$consine(X,Y)=\frac{\sum_{i=1}^{N}x_i*y_i}{\sqrt{\sum_{i=1}^{N}x_i^2}\times\sqrt{\sum_{i=1}^{N}y_i^2}} \quad (3)$$

则查询向量 q 中研究人员查询内容 q_r 与图模型中的各个文章的内容相似度向量 q_p 可表示为

$$q_p(q_r)=\left[consine(q_r,pv_1),consine(q_r,pv_2),\ldots\ldots,consine(q_r,pv_{k-1}),consine(q_r,pv_k)\right] \quad (4)$$

根据上述过程可以得到查询向量 q_p 的 distributed representation 表示。

c）在关键字层，本文利用上述 word2vec 的方法生成关键字的向量，关键字 w_i，w_j 的向量用 **X** 和 **Y** 表示。而词与词之间的关系则用两者之间的 consine 相似度来表示。计算公式如公式 3 所示，则有 q_w（w_i，w_j）=consine（X,Y）。

最终查询向量为 $q=[q_p,\ q_a,\ q_w]$，使用上述三层图模型，利用上面的概率转移矩阵 **M** 运行推荐模型。

1.3 推荐模型

为了能给目标文献推荐出合适的文章，需要计算目标文章和候选文章之间的相关性大小。因此，本文用到了重启动随机游走（Random Walk with Restarts，RWR）算法[18]。用 $R(v_i)$ 表示图 **G** 中的节点与目标文献的相关性。**G** 中所有节点与目标文献的相关性向量 **R** 可以通过幂迭代得到，计算公式如下所示：

$$R^{(t+1)}=(1-\theta)MR^{t}+\theta q \quad (5)$$

其中：θ 是返回起始节点的重启概率，**M** 是图 **G** 的权重矩阵。在公式 5 中，初始时令 $R^{(0)}=q$。计算出目标文献与候选文献的相关性 **R** 后，本文所要的结果是对论文的评分值，即 $R(p)$，选取前 k 篇 **R** 值较大的论文作为推荐返回。为使公式 5 在经过若干次幂迭代之后能够收敛，需要对 **M** 和 q 进行列归一化。

本文的推荐算法过程如下：

a)数据准备。对数据集中的论文内容进行提取，并对数据集中作者—文章关系、作者—作者关系、文章—文章关系、文章—关键字关系进行提取。

b)模型构建。构建多关系图模型，用矩阵 **M** 表示，利用 word2vec 生成词向量，并根据词向量生成文本向量，计算词与

词之间的相似度，并计算研究人员输入的查询内容与候选文章之间的关联性，并根据研究人员生成查询作者向量；

c)循环迭代。依照式（5）所示的随机游走进行计算，更新 ***R*** 并与 ***M*** 相乘不断迭代，直到 ***R*** 收敛，得到 ***R*** 的最终结果。

d)结果输出。对 ***R*** 最终值，选取对应的文章维度向量，并排序，输出文章的 top*N* 的列表。

2　实验与分析

为了验证本文中所提的算法的有效性，本小节将进行文中所提出的方法在召回率（recall）和归一化折损累积增益（NDCG）这两个评价指标与其他不同的方法做实验对比。并介绍了常用的一个数据集合 AAN（Association of Computational Linguistics (ACL) anthology network）[19]。

2.1　实验数据

本文所做的验证测试均是在 AAN 数据集上进行的，该数据集包含了许多 ACL 期刊上的所有论文。未经处理的数据集合包含从 1965 年到 2013 年的 21，236 篇文章，并且有论文内容包括摘要和题目，论文出版年份，作者和期刊等信息。在对文章关键词的表示时，本文用论文标题和摘要来表示，由于数据集合中并没有给出文章的摘要信息，需要对摘要进行提取工作，并生成关键词。先通过下面的方法对文章内容进行预处理：

a)删除没有引用关系的论文；

b)提取文章摘要和标题，并删除没有摘要和标题的论文；

c)删除由少于三个字符组成的单词；

d)删除停用词；

e)用 NLTK 的 stemmer 词干分析工具对单词进行词干化，为了减少噪声数据，还删除了在整个数据集的语料中词频少于十次的单词，这样一共产生的 4，918 个不重复的单词。

本文使用 2013 年之前发表的 11 129 篇论文候选文章集合来建立多关系的图模型，并把 2013 年发表的 1,375 篇论文的作者身份 ID 作为查询向量的作者部分，关键词与候选文章关键词的相似度作为查询向量的内容部分，查询内容与候选文章内容的相似度作为查询向量的文章部分。对这 1 375 篇论文进行引用文献推荐。表 2 给出了本文用到的 AAN 数据集进行处理之后的基本统计信息。被引用的文章表示在相应的范围内至少被引用了一次的文章，引用关系表示在相应的范围内总共的引用关系。

表 2 AAN 数据集的基本统计信息

年份	论文数目	作者数目	被引用的文章	引用关系
2013 年之前	11,129	9, 744	9, 016	65, 891
2013 年	1,375	1, 333	4, 822	11, 529

2.2　评价指标

本文用 recall@*N* 值[20]和 NDCG@*N* 值[3]进行来对推荐结果的准确性和排序质量评测，这两种方法被广泛应用于信息检索和统计分类领域。

a)recall@*N*。在信息检索领域，recall@*N* 衡量的是检索系统的查全率。在本文中 recall@*N* 是 Top-N 的推荐列表中实际被引用的文献数量与目标文献实际引用列表中文献数量的比值。计算公式可以表示为

$$recall@N=\frac{1}{C}\sum_{i=1}^{C}\frac{R(p)\cap T(p)}{T(p)} \qquad (6)$$

其中：C 表示的是查询列表的个数，*N* 表示的是推荐列表的个数。*T*(*p*)表示目标文献实际引用的文献集合，*R*(*p*)表示推荐结果中文献集合。所以 *R*（*p*）∩T（*p*）表示推荐列表中实际引用的文献集合。

b)NDCG@*N*。recall@*N* 并不能充分评估推荐方法的有效性，一个好的推荐系统对实际相关的引用文献在推荐结果中的位置应该是敏感的，显然，recall@*N* 并不具有这样的功能。本文希望推荐结果中的相关文献出现在推荐列表的靠前位置，因此在这篇文章中用到了 *NDCG@N* 来衡量推荐列表的排序。NDCG@*N* 的定义如下：

$$NDCG@N=\frac{1}{C}\sum_{i=1}^{C}\left\{\left\{\sum_{j=1}^{N}\frac{2^{r_j}-1}{\log_2(j+1)}\right\}/IDCG@N\right\} \qquad (7)$$

其中：C 和 *N* 的表示和上文中 *recall@N* 中的表述相同，r_j 表示推荐结果列表中，排名为 *j* 的文献的评级，$r_j\in\{0,1\}$ss，$r_j=1$ 表示该文章是一个相关文献，$r_j=0$ 则意味着不相关。*IDCG@N* 则是理想状态下的推荐结果排序，计算公式如下：

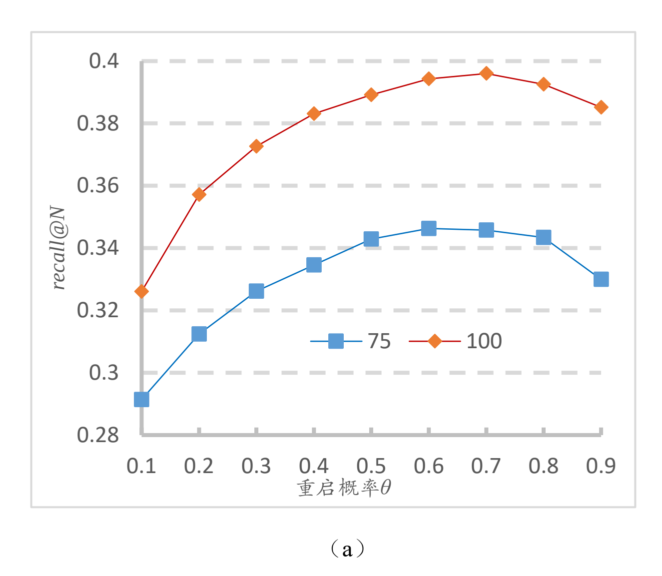
$$IDCG@N=\sum_{i=1}^{|REL|}\frac{2^{r_i}-1}{\log_2(i+1)} \qquad (8)$$

其中：*REL* 表示推荐结果中实际相关的论文集合。

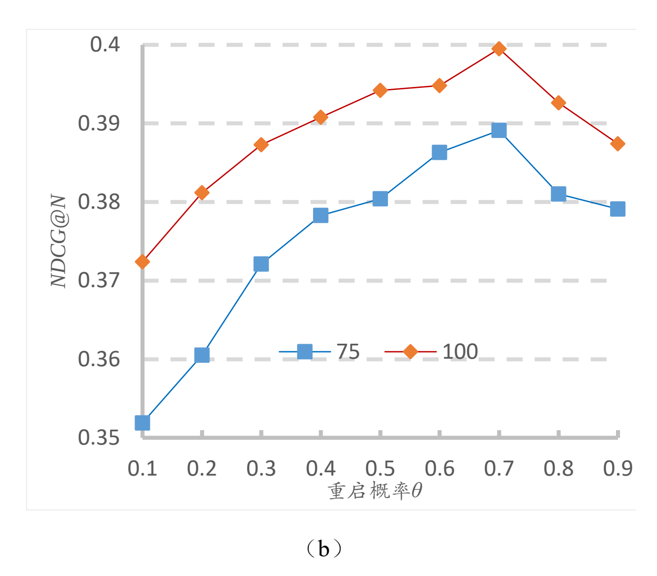
2.3　参数调整与分析

在推荐系统中，最初的概率转移矩阵，通过迭代过程中不断重启动随机游走模型最终确定节点的概率转移矩阵。本节中，将主要分析一个 1.3 小节中所提到的重启概率参数 θ。

不同的 θ 值对最终的推荐结果的质量有着不同的影响，所以本小节中通过本文提出的方法使用不同的 θ 执行相关实验。对一个查询向量的节点来说，（1-θ）表示从当前节点过渡到相邻节点的概率，而 θ 则代表了从当前节点过渡到初始查询向量（查询论文）的起始节点的概率。公式 5 中的 θ 的值越大意味着返回初始查询向量中的节点的概率越大。图 2 显示了 θ 从 0.1 到 0.9 时的 *recall@75*、*recall@100* 和 *NDCG@75*、*NDCG@100* 的实验对比。



（a）



（b）

图 2 不同重启概率 θ 的 *recall* 和 *NDCG* 的值的变化。（a）图为 *recall@75* 和 *recall@100*，（b）图为 *NDCG@75* 和 *NDCG@100*。

如图 2（a）所示，在 0.1~0.6 内，随着 θ 值的增大，*recall@75* 和 *recall@100* 随之变大。当 θ 从 0.6 改变成 0.7 时，*recall@75* 减小，而 *recall@100* 则变大，但 *recall@75* 减小幅度不大，当 θ 大于 0.7 时这两个评测指标都开始下降。从而本文得出，当 θ=0.7 时，*recall@75* 和 *recall1@100* 达到最优值，并且从图中观察到，当 θ=0.1 时，*recall*@75 和 *recall@100* 的结果值最差。

如图 2（b）所示，*NDCG@75* 和 *NDCG@100* 的变化曲线大致相同，当 θ 在 [0.1,0.7]间时，随着 θ 值的变大，*NDCG@75* 和 *NDCG@100* 总体呈上升趋势，在[0.7,0.9]之间，这两个测量指标呈下降趋势。由此得出，在 θ=0.7 时，NDCG@75 和 NDCG@100 的值达到最大，当 θ=0.1 时，*NDCG@75* 和 *NDCG@100* 结果最小。

根据图 2 可以得到当 θ=0.7 时，*recall@N* 和 *NDCG@N* 达到最大值，因此本文把 θ 的值确定为 0.7。根据相关分析表明不同的重启概率 θ 对本文所提方法的推荐结果有着不同的影响，并且当设置 θ=0.7 时，本文提出的方法的推荐效果达到最好。

2.4 对比实验

通过五个不同方法和本文提出的方法作对比来说明本文提出方法的有效性。这五种方法如下所示：

a)关联主题模型[21]（**relational topic model，RTM**）。RTM 是 sLDA 的一个扩展算法，用链接作为监督来训练 LDA 模型。RTM 在两个不同的文本数据之间增加了一个二元随机变量，利用主题分布的带有 Hadamard 积的 SIGMOD 函数来推荐论文。在本文实验中，设置 RTM 的主题数目为 60。

b)Link-PLSA-LDA[22]。Link-PLSA-LDA 使用 PLSA 生成引用博客和 LDA 生成引用的博客，在本文中，把每篇文章当成一个博客，推荐列表是根据引文中论文的主题-词分布的相关性和引用论文中论文分布的相关性生成的。在本实验中，Link-PLSA-LDA 主题数目同样设置为 60。

c)LDA。LDA 是一个文档主题生成模型，包含文档、词、主题三个部分。在本实验中首先使用 LDA 得到主题信息，之后推荐与查询高度相关的主题下的文章，主题数目设置为 70。

d)CiteRank[23]。CiteRank 为每一个节点使用个性化的传送因素，并且考虑到了时间衰减，使用交通动力学的一个简单的网络模型来对科技论文进行排序，CiteRank 使用到了统计力学，通信供应和信息网络的知识内容。在本实验中，该方法的衰减参数 τ 设置为 2.6。

e)PopRank[24]。PopRank 是 PageRank 的一种扩展算法。该方法将一个流行的传播因子（PPF）添加到一个对象的每一个引用，并且利用作者论文关系和出版信息对候选论文进行排序。本实验中设置流行的传播因子 PPF=0.3，阈值为 0.01。

本文提出了一种基于 RandomWalk 的论文、作者、论文关键字的三层图模型的方法（PAWRW），查询中使用到了论文相似度，作者关系，关键字相似度作为的查询向量，并且考虑了不同层之间的转换概率。

表 3 是不同的方法性能对比的结果，明显地可以看到随着 *N* 的增大，*recall@N* 和 *NDCG@N* 的值都会随之变大。从不同的方法对比中可以看出，PAWRW 的实验结果均高于 RTM、Link-PLSA-LDA、LDA、CiteRank、PopRank 这五种方法在 AAN 数据集上，这是因为本文的实验模型中融入了内容和网络信息还有作者信息，而仅仅通过引用关系或者是内容信息来进行相关论文的推荐具有一定的局限性。由于缺少文本信息，从表 3 中可以发现在所有指标上 PopRank 的效果明显比其他方法要差。由于仅仅依赖文本主题相似度信息，LDA 各项指标的结果是所有方法中最差的。RTM 和 Link-PLSA-LDA 的实验结果较为相近，并且要优于 LDA，这是因为这两种方法整合了引用链接信息和其他一些额外的信息来进行主题学习。通过表 3 还可以看出，CiteRank 与 LDA 相比，即在论文内容的基础上使考虑了时间因素推荐结果并没有得到明显提高。

表 3 不同方法的对比结果

Top-N	25		50		75		100	
方法	*recall*	*NDCG*	*recall*	*NDCG*	*recall*	*NDCG*	*recall*	*NDCG*
PAWRW	0.2176	0.3084	0.2853	0.3429	0.3458	0.3891	0.396	0.3995
RTM	0.1734	0.2738	0.2751	0.3225	0.3273	0.3851	0.3698	0.3841
Link-PLSA-LDA	0.1725	0.2742	0.2576	0.3174	0.3207	0.3815	0.3647	0.3786
PopRank	0.1341	0.1124	0.2254	0.2149	0.2911	0.2635	0.3092	0.2782
LDA	0.1132	0.037	0.1473	0.083	0.1755	0.1207	0.1826	0.1865
CiteRank	0.1233	0.052	0.1506	0.9254	0.1783	0.1327	0.1907	0.1941

2.5 不同查询向量的对比分析

最后，为了分析查询向量对 PAWRW 推荐结果的影响，本小节将对比不同向量的推荐结果，q_1 表示查询向量中只有作者信息，没有论文相似度信息和关键字信息，即 $q_1=[0, q_a, 0]$；$q_2=[q_p, q_a, 0]$，包含作者信息和论文相似度信息，而不包括关键字信息；$q_3=[0, q_a, q_w]$，包含作者信息和关键字信息，但不包括文本相似度信息；q_4 则是本文中所用查询，包含了作者信息，论文相似度信息，关键字信息，$q_4=[q_p, q_a, q_w]$。表 4 为各不同查询向量的推荐结果。

通过表 4 可知，在查询语句中添加文章相似度信息和关键字的信息都可以使推荐效果得到提升，并且发现，添加关键字信息之后的推荐效果要比添加文章相似度信息的推荐效果好。同时，还可以发现，查询为 q_3 和 q_4 情况下的推荐效果比较接近，特别是 *recall@N* 的值。这是因为在论文相似度的计算过程中，虽然也用到了 word2vec，但比关键字相似度的计算多了一步，就是文本向量进行了简单的相加，使本来就有误差的词向量生成的文本向量误差更大，所以查询 q_3 的推荐效果要比 q_2 的推荐效果高，同时也从侧面说明了查询 q_4 的推荐效果为什么与 q_3 的推荐效果在性能指标上较为接近。

表 4 不同查询向量推荐结果对比

Top-N	25		50		75		100	
	recall	*NDCG*	*recall*	*NDCG*	*recall*	*NDCG*	*recall*	*NDCG*
q_1	0.1446	0.1846	0.2171	0.3207	0.2617	0.3518	0.299	0.3541
q_2	0.1565	0.2537	0.2363	0.3469	0.2844	0.3812	0.3213	0.3885
q_3	0.1749	0.2743	0.2722	0.3128	0.3384	0.3402	0.3931	0.3716
q_4	0.2176	0.3084	0.2853	0.3429	0.3458	0.3891	0.396	0.3995

3 结束语

本文提出了一种结合论文引用关系、作者关系、论文内容等信息的三层图模型的文本推荐方法。考虑到一篇论文被推荐不仅仅与同时被推荐的论文相关，还与这篇论文的作者和该论文的内容相关，并在不同层次类型的关系对象上使用不同的参数，使某节点在向不同层次的节点跳转时的概率不同，构建概率转移矩阵，针对查询向量中的文本信息缺失问题，提出了利用 word2vec 的方法计算文本相似度及词的相似度对查询向量进行填充，有效利用了文本内容的上下文信息。之后基于 Random Walk 对相应的查询给出相关的推荐结果。实验表明，改进的查询向量和基于三层图模型的推荐方法可以提高个性化引文推荐的效果，突出相关论文的排序结果。本文并没有考虑模型中词关系的信息，未来的研究工作在对模型中词关系进行下一步的研究和优化的同时，建立更有效的推荐模型。

参考文献：

[1] Ren Xiang, Liu Jialu, Yu Xiao, *et al.* Cluscite: effective citation recommendation by information network-based clustering [C]// Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014: 821-830.

[2] 李伟霖，王成良，文俊浩. 基于评论与评分的协同过滤算法 [J]. 计算机应用研究, 2017, 34 (2): 361-364. (Li Weilin, Wang Chengliang, Wen Junhao. Collaborative filtering recommendation algorithm based on reviews and ratings [J]. Application Research of Computers, 2017, 34 (2): 361-364.)

[3] Totti LC, Mitra P, Ouzzani M, *et al.* A query-oriented approach for relevance in citation networks [C]// Proc of the 25th International Conference Companion on World Wide Web. 2016: 401-406

[4] 王宁宁，鲁燃，王智昊，等. 基于用户标签的微博推荐算法 [J]. 计算机应用研究, 2017, 34 (1): 58-61. (Wang Ningning, Lu Ran, Wang Zhihao, *et al.* Microblog recommendation algorithm based on user’s tag [J]. Application Research of Computers, 2017, 34 (1): 58-61.)

[5] 吴一帆，王浩然. 结合用户背景信息的协同过滤推荐算法 [J]，计算机应用, 2008, 28 (11): 2972-2974. (Wu Yifan, Wang Haoran. Collaborative filtering algorithm using user background information [J]. Computer Applications, 2008, 28 (11): 2972-2974.)

[6] 王茜，王均波．一种改进的协同过滤推荐算法 [J]. 计算机科学, 2010, 37 (6): 226-228. (Wang Qian, Wang Junbo. Improved Col laborative Filtering Recommendation Algorithm [J]. Computer Science, 37 (6): 226-228.)

[7] Yang Zhe, Wu Bing, Zheng Kan, *et al.* A survey of collaborative filtering-based recommender systems for mobile Internet applications [J]. IEEE Access, 2016, 4: 3273-3287

[8] Xia Feng, Wang Wei, Bekele T M, *et al.* Big scholarly data: a survey [J]. IEEE Trans on Big Data, 2017, 3 (1): 18-35

[9] Pan Linlin, Dai Xinyu, Huang Shujian, *et al.* Academic paper recommendation based on heterogeneous graph [M]. [S. l.] : Springer International Publishing, 2015: 81–392

[10] West J D, Wesleysmith I, Bergstrom C T. A recommendation system based on hierarchical clustering of an article-level citation network [J]. IEEE Trans on Big Data, 2016, 2 (2): 113-123

[11] Zhou Ding, Zhu Shenghuo, Yu Kai, *et al.* Learning multiple graphs for document recommendations [C]// Proc of the 17th International Conference on World Wide Web. New York: ACM Press, 2008: 141-150

[12] Meng Fanqi, Gao Dehong, Li Wenjie, *et al.* A unified graph model for personalized query-oriented reference paper recommendation [C]// Proc of ACM International Conference on Conference on Information & Knowledge Management. New York: ACM Press, 2013: 1509-1512.

[13] 郭蓝天，李扬，慕德俊，等．基于 LDA 主题模型的话题发现方法 [J]. 西北工业大学学报, 2016, 34 (4): 698-702. (Guo Lantian, Li Yang, Mu Dejun *et al.* A LDA model based topic detection method [J]. Journal of Northwestern Polytechnical University, 2016, 34 (4): 698-702)

[14] 许珂，蒙祖强，林启峰．基于语义关联和信息增益的 TFIDF 改进算法研究 [J]. 计算机应用研究, 2012, 29 (2): 557-560. (Xu Ke, Meng Zuqiang, Lin Qifeng. Improved TFIDF feature extraction algorithm based on semantic association and inflormatian gain [J]. Application Research of Computers, 2012, 29 (2): 557-560.)

[15] Wang Chong, Blei D M. Collaborative topic modeling for recommending scientific articles [C]// Proc of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Minin g. New York: ACM Press, 2011: 448-456.

[16] Li Jing, Xia Feng, Wang Wei, *et al.* Acrec: a co-authorship based random walk model for academic collaboration recommendation [C]// Proc of the 23rd International Conference on World Wide Web. New York: ACM Press, 2014: 1209-1214

[17] Mikolov T, Chen Kai, Corrado G, *et al.* Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv: 13013781, 2013

[18] 黄斌．社会网络中基于随机游走的名称消歧算法 [J]. 计算机应用研究, 2015, 32 (12): 3650-3653 (Huang Bin. Random walk based name disambiguation algorithm in social networks [J]. Application Research of Computers, 2015, 32 (12): 3650-3653.)

[19] Radev D R, Muthukrishnan P, Qazvinian V, *et al.* The ACL anthology network corpus [J]. Language Resources and Evaluation, 2013, 47 (4): 919-944

[20] Liu Qi, Chen Enhong, Xiong Hui, *et al.* Enhancing collaborative filtering by user interest expansion via personalized ranking [J]. IEEE Trans on Systems, Man, and Cybernetics, Part B (Cybernetics) , 2012, 42 (1): 218-233

[21] Chang Jonathan, Blei D M. Relational topic models for document networks [C]// Proc of the 12th International Conference on Artificial Intelligence and Statistics. 200: 81-88.

[22] Nallapati R M. , Ahmed A, Xing E P, *et al.* Joint latent topic models for text and citations [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 542-550.

[23] Walker D, Xie Huafeng, Yan K K, *et al.* Ranking scientific publications using a simple model of network traffic [J]. Journal of Statistical Mechanics-Theory and Experiment, 2006, 6 (6): P06010-1 – P06010-10.

[24] Nie Zaiqing, Zhang Yuanzhi, Wen Ji Rong, *et al.* Object-level ranking: bringing order to Web objects [C]// Proc of International Conference on World Wide Web. New York: ACM Press, 2005: 567-574